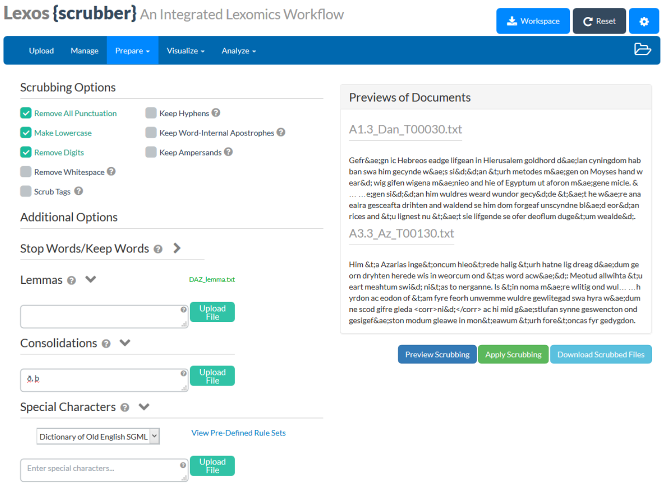
Lexos: An Integrated Lexomics Workflow

Scott Kleinman
scott.kleinman@csun.edu
California State University: Northridge, United States

Mark LeBlanc
mleblanc@wheatoncollege.edu
Wheaton College, United States

Lexos is a browser-based suite of tools that helps lower barriers of entry to computational text analysis for humanities scholars and students. Situated within a clean and simple interface, *Lexos* consolidates the common pre-processing operations needed for subsequent analysis, either with *Lexos* or with external tools. It is especially useful for scholars who wish to engage in research involving computational text analysis and/or wish to teach their students how to do so but lack the time for a manual preparation of texts, the skill sets needed to prepare their texts analysis, or the intellectual contexts for situating computational methods within their work. *Lexos* is also targeted at researchers studying early texts and texts in non-Western languages, which may involve specialized processing rules. It is thus designed to facilitate advanced research in these fields even for users more familiar with computational techniques. *Lexos* is developed by the Lexomics research group led by Michael Drout (Wheaton College), Mark LeBlanc (Wheaton College), and Scott Kleinman (California State University, Northridge). It is built on Python 2.7-Flask microframework, with jQuery-Bootstrap UI, and visualizations in d3.js. The Lexomics research group provides access to an public installation of *Lexos* which does not retain data after a session has expired. Users may also install *Lexos* locally by cloning the GitHub repository.

Lexos guides users through a workflow of steps that reflects effective practices when working with digitized texts. The workflow includes: (i) uploading Unicode-encoded texts in plain text, HTML, or XML formats; (ii) "scrubbing" functions for consolidating preprocessing decisions such as the handling of punctuation, white-space, and stop words, the use of lemmatization rules, and the handling of embedded markup tags and special character entities; (iii) "cutting" texts into segments based on the number of characters, tokens, or lines, or by embedded milestones such as chapter breaks; (iv) tokenization into a Document Term Matrix of raw or proportional counts using character or word n-grams; (v) visualizations such as comparative word clouds per segment (including the ability to visualize topic models generated by MALLET); Rolling Window Analysis that plots the frequency of string, phrase, or regular expression patterns or pattern-pair ratios over the course of a document or collection; and (vi) analysis tools including statistical summaries, hierarchical and k-means clustering, cosine similarity rankings, and Z-tests to identify the relative prominence of terms in documents, document classes, and the collection as whole. At each stage in the workflow the user may download data, visualizations, or the results of the analytical tools, along with metadata about their preprocessing decisions or the parameters selected for their experiments. *Lexos* thus enables the export of data for use with other tools and facilitates experimental reproducibility.



Scrubbing Options

☑ Remove All Punctuation
☑ Make Lowercase
☑ Remove Digits
☐ Remove Whitespace
☐ Scrub Tags
☐ Keep Hyphens
☐ Keep Word-Internal Apostrophes
☐ Keep Ampersands

Additional Options

Stop Words/Keep Words

Lemmas: DAZ_lemma.txt

Consolidations: ð,þ

Special Characters: Dictionary of Old English SGML | View Pre-Defined Rule Sets

Enter special characters...

Previews of Documents

A1.3_Dan_T00030.txt

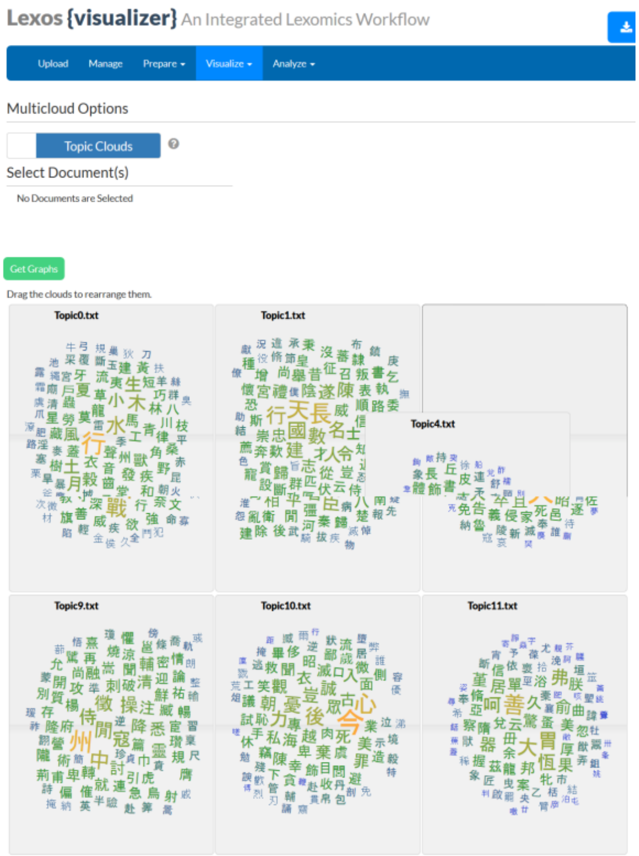
Gefr&ae;gn ic Hebreos eadge lifgean in Hierusalem goldhord d&ae;lan cyningdom hab ban swa him gecynde w&ae;s si&d;&d;an &t;urh metodes m&ae;gen on Moyses hand w ear&d; wig gifen wigena m&ae;nieo and hie of Egyptum ut aforon m&ae;gene micle. & … …e;gen si&d;&d;an him wuldres weard wundor gecy&d;de &t;&ae;t he w&ae;re ana ealra gesceafta drihten and waldend se him dom forgeaf unscyndne bl&ae;d eor&d;an rices and &t;u lignest nu &t;&ae;t sie lifgende se ofer deoflum duge&t;um wealde&d;.

A3.3_Az_T00130.txt

Him &t;a Azarias inge&t;oncum hleo&t;rede halig &t;urh hatne lig dreag d&ae;dum ge orn dryhten herede wis in weorcum ond &t;as word acw&ae;&d;: Meotud allwihta &t;u eart meahtum swi&d; ni&t;as to nerganne. Is &t;in noma m&ae;re wlitig ond wul… …h yrdon ac eodon of &t;am fyre feorh unwemme wuldre gewlitegad swa hyra w&ae;dum ne scod gifre gleda <corr>ni&d;</corr> ac hi mid g&ae;stlufan synne geswencton ond gesigef&ae;ston modum gleawe in mon&t;eawum &t;urh fore&t;oncas fyr gedygdon.

Figure 1: The *Lexos* Scrubber Tool

Lexos addresses three significant challenges for our intended users. The first challenge involves the **adoption** of computational text analysis methods. Many approaches require proficiency with command line scripting or the use of complex user interfaces that require time to master. *Lexos* addresses this problem through a simple, browser-based interface that manages workflow through the three major steps of text analysis: pre-processing, generation of statistical data, and visualization. In this, *Lexos* resembles *Voyant Tools* (Sinclair and Rockwell, 2016), although *Lexos* places more emphasis on and providing more tools for preprocessing and segmenting texts. *Lexos* also shares with tools like *Stylometry with R* (Eder, et al., 2013; Eder, 2013) and emphasis on cluster analysis, providing both hierarchical and K-Means clustering with silhouette scores as limited form of statistical validation. While *Lexos* is not a topic modeling tool, it provides a useful “topic cloud” feature for MALLET data that will be useful for beginners since there are few accessible ways to visualize MALLET output that work well out of the box.



No Documents are Selected

Drag the clouds to rearrange them.

Figure 2: The *Lexos* Multicloud tool showing Chinese "topic clouds"

The second challenge is the **opacity** of the procedures required to move between computational and traditional forms of text analysis. In order to reduce the “black boxiness” of algorithmic methods, *Lexos* contains an embedded component called “In the Margins” which provides non-technical explanations of the statistical methods used and effective practices for handling situations typical of humanities data. “In the Margins” is a Scalar “book” which can be read separately; however, its individual pages are embedded in *Lexos* using Scalar’s API, making them easily accessible for users of the tool. *Lexos* shares with tools like *Voyant* an engagement with the hermeneutics of text analysis and attempts to embed “In the Margins” discussion of these issues in the user interface close to the user’s workflow. We hope “In the Margins” will host advice and commentary from contributors with the Digital Humanities community.

A third challenge is the **tension** between quantitative and computational approaches and the traditions of theoretical and cultural criticism that dominate the humanities in the academy. As Alan Liu (2013) has recently argued, the challenge is to give a better theoretical grounding to the hybrid quantitative-qualitative method of the Digital Humanities by exploring the ways in which we negotiate the difficulties imposed by “the aporia between tabula rasa quantitative interpretation and humanly meaningful qualitative interpretation” (414). The design of *Lexos* and the discussions in “In the Margins” are intended to open a space for discussion of issues related to the opacity of algorithmic approaches and the limitations and epistemological challenges of computational stylistic analysis and visual representation of humanities data.

This poster presentation provides demonstrations of *Lexos* using some literature from Old, Middle, and Modern English, as well Chinese, which are in our current test suite. We also discuss use cases and best practices, how to install *Lexos* locally, and how scholars may contribute to the still growing content of “In the Margins”.

Bibliography

Drout, M., Kleinman, S., and LeBlanc, M. 2016-. “In the Margins.” http://scalar.usc.edu/works/lexos./

Eder, M. (2013). “Mind Your Corpus: Systematic Errors in Authorship Attribution.” *Literary and Linguistic Computing* 28 (4): 603–14.

Eder, M., Kestemont, M., and Rybiki, J. 2013. “Stylometry with R: A Suite of Tools (Abstract of Poster Session)”. Presented at Digital Humanities 2013, Lincoln, Nebraska. http://dh2013.unl.edu/abstracts/ab-136.html, https://sites.google.com/site/computationalstylistics/

Kleinman, S., LeBlanc, M.D., Drout, M. and Zhang, C. 2016. *Lexos* v3.0. https://github.com/WheatonCS/Lexos/.

Liu, A. (2013). “The Meaning of the Digital Humanities.” *PMLA* 128 (2): 409-23.

McCallum, A.K. (2002). *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu.

Sinclair, S., and Rockwell, G. (2016). *Voyant Tools*. Web. http://voyant-tools.org/.